Capstone Project-1 Summary

Organizations are facing increased supply chain costs and competitive pressures. Caterpillar, who manufactures construction equipment requires 8,855 different styles of tube assemblies supplied by 57 different suppliers. Some suppliers carry just one type of assembly while others carry many depending upon their length of the contract. Furthermore, slight difference in specifications such as bend radius, length, weight, order quantity and so on, the tube assembly price can vary and so it becomes difficult to identify which assemblies we buy in bulk and from whom so as to realize best business savings?

Hence, the primary goal of this project was to be able to gather actionable business insights so that we can optimize supply chain performance and improve operational sustainability.

Here are the two basic questions we will try to answer:

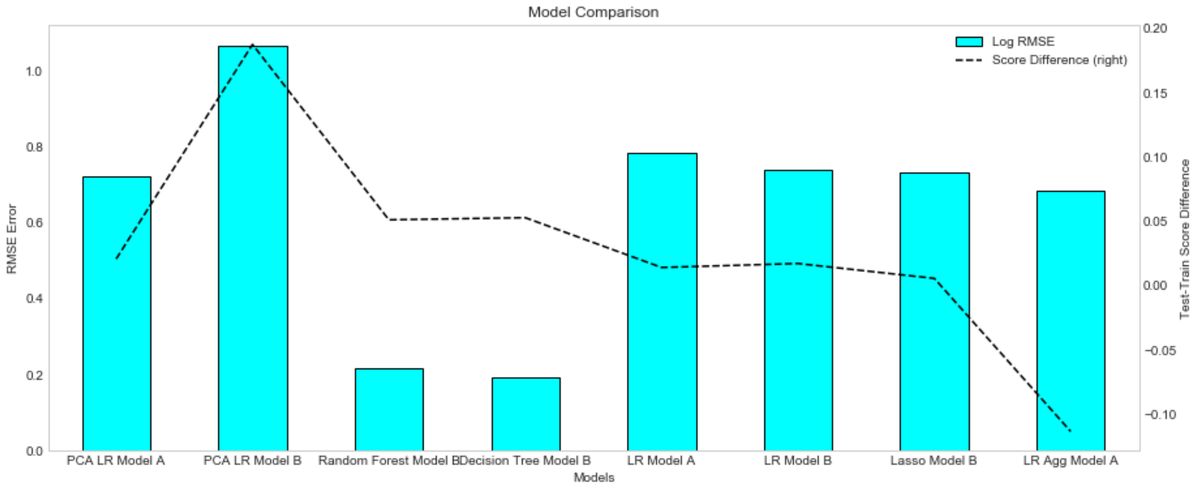
1. Would it possible to build a model that can learn from our previous organizational spend and help us predict supplier pricing based on different specification?
2. Because, we have so many different suppliers and assemblies, would it be possible to categorize and see which assemblies and suppliers best meets our business needs?

My first step in this project started with gathering and qualifying public datasets. I chose Caterpillar because I felt the connection. There were 21 tables which had to be combined to preserve the most relevant features for our modelling. We wrangled the data, developed new features as well as dropped fields to make the most out of 23% of the entire dataset.

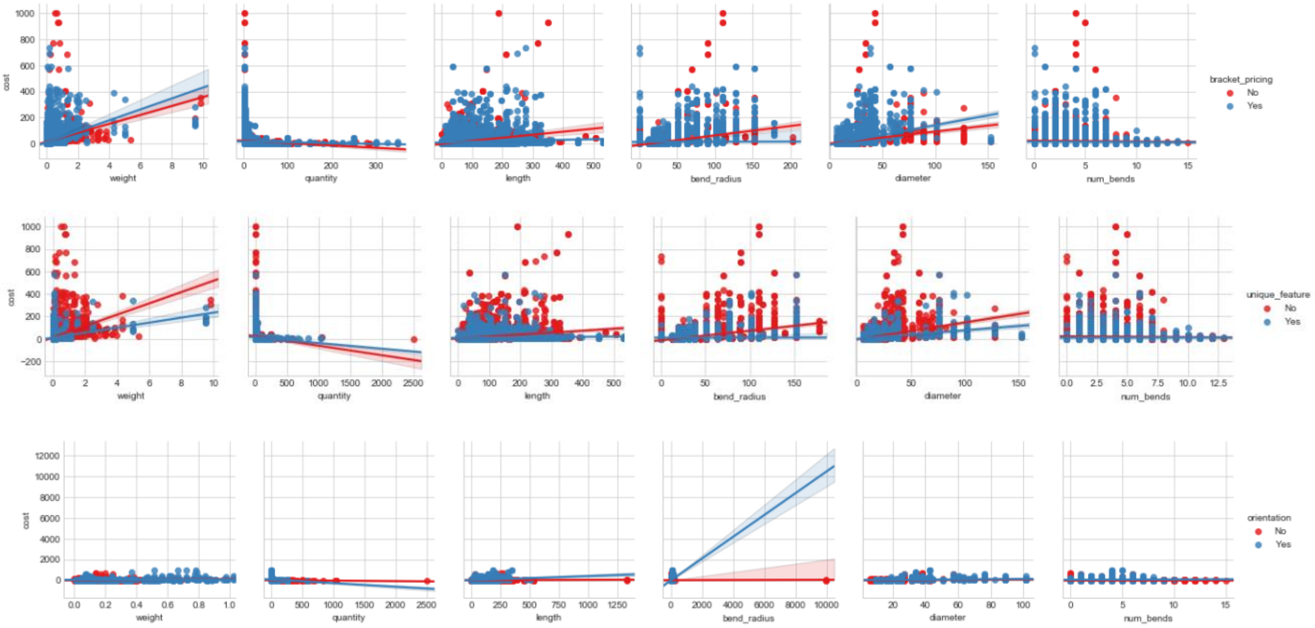
Please see below highlights of my project findings. I have used Jupyter Notebook to ask questions and prepare the code. You will observe visuals as well as observations along the way which has helped me unfold hidden insights.

Section1: Supplier Pricing Prediction:

1) Model Prediction: Just to get started, I used few different models to compare test-train scores and rmse error. I noticed that ensemble algorithm (Random Forest) resulted in best accuracy 98% and optimized rmse error on training and validation test set.

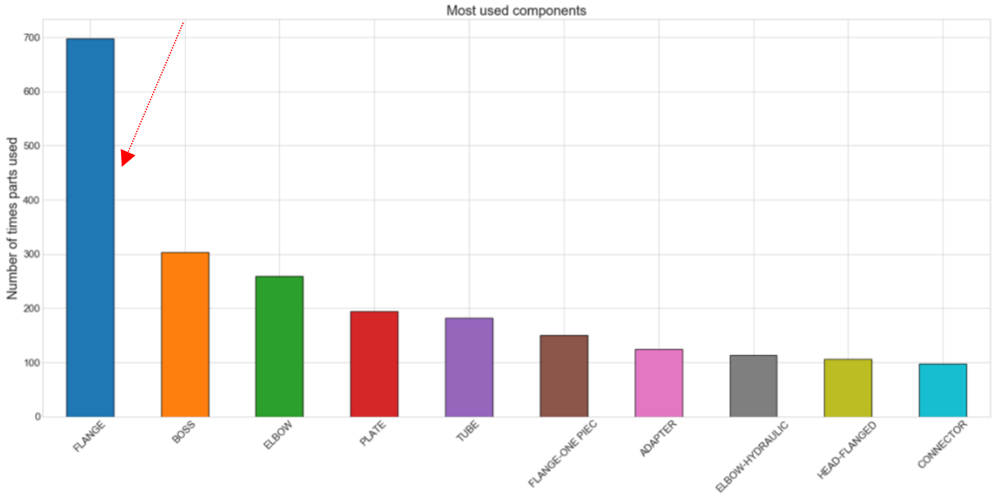


2) Supplier pricing was found dependent on product specifications, annual volume and type of contract pricing offered. Cost Vs Weight/Length/Dimeter etc.

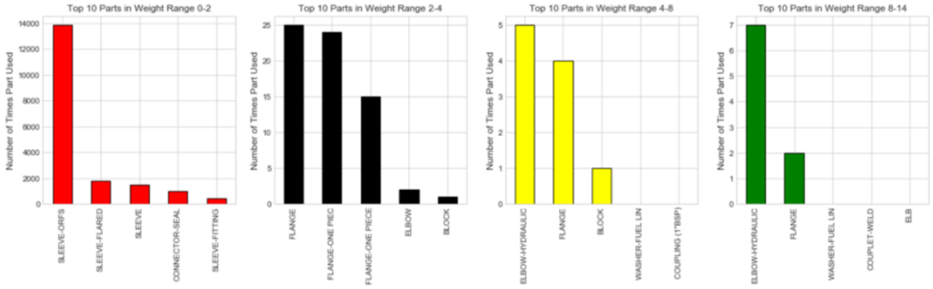


3) Because cost was dependent on the weight of the assembly, it was important to identify which components were used frequently and what was their weight distribution. This can be very useful especially when business wants to compare total cost across multiple suppliers.

Overall Distributuion of Most Used Components:



Most Used Components by Different Weight Range:



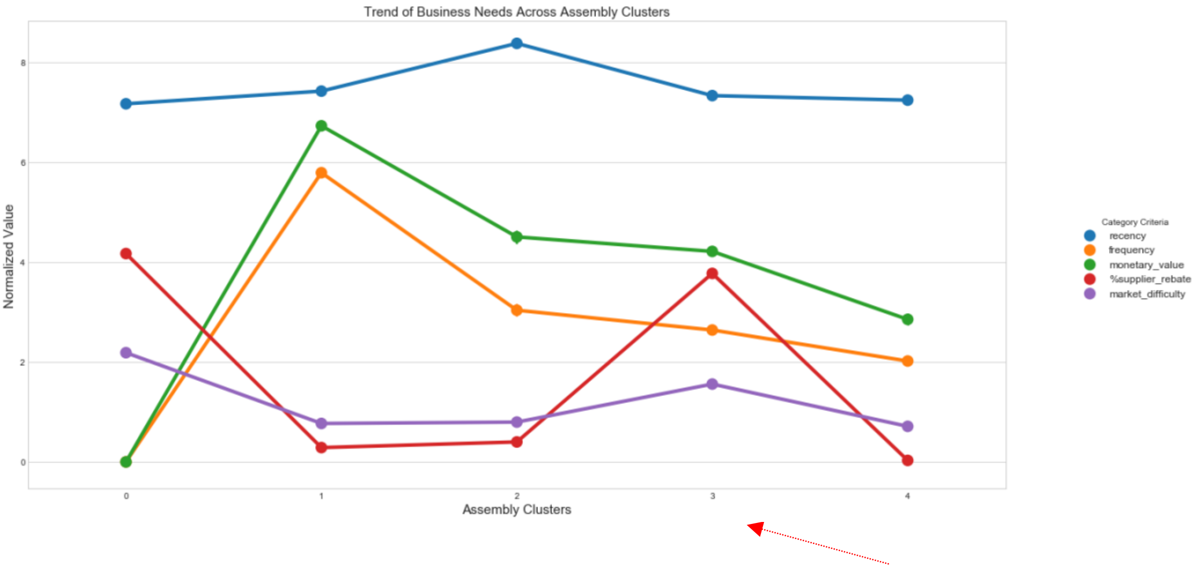
Section 2: Assembly and Supplier Categorization:

1) Pre-Classification: It was important to establish a pre-categorization baseline such as contract length, monetary spend, supplier preference etc. to understand and assess assembly and supplier trends.

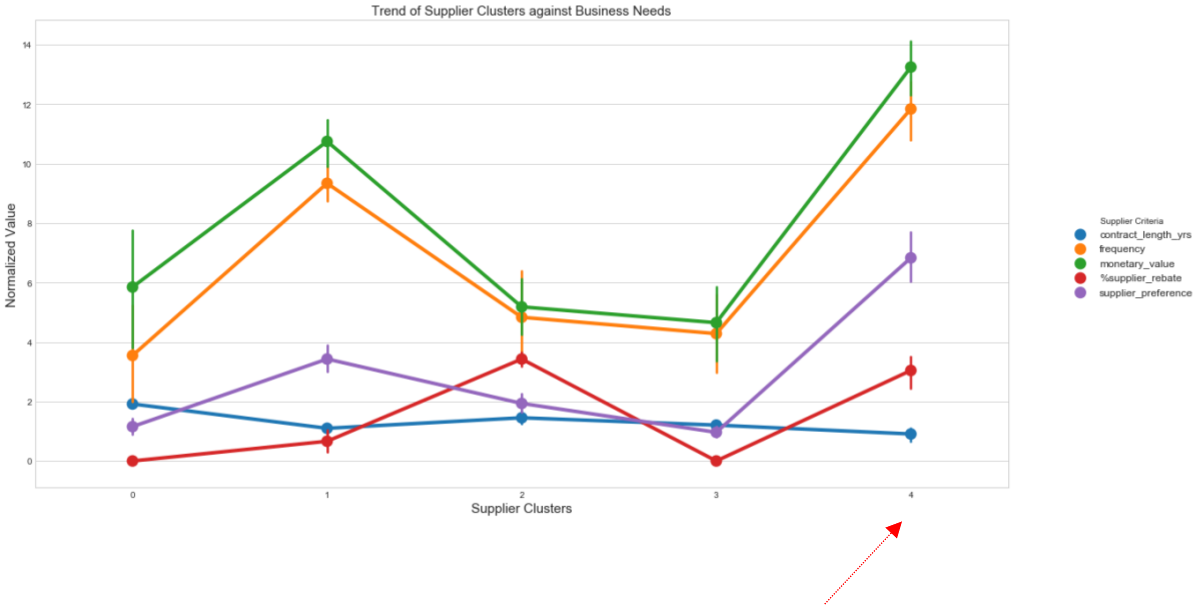
2) Clustering: Then, apply algorithm to group suppliers and assemblies using pre-set criteria.

3) Analysis: Identify groups of assemblies and suppliers that can be benchmarked against business needs and can be utilized in learning and improving efficiency of the other clusters.

a) Best Managed Assembly Category: Cluster 3



b) Best Managed Suppliers: Supplier Cluster 4



Future Possibilities:

1) Add additional features such as on time delivery, safety performance, contract compliance and inventory to understand other supplier performance factors.
2) Develop ETL pipeline to see and optimize supply chain performance in real-time.
3) Use time series modelling to predict future supplier pricing.

For more information, please refer to my notebook on github.

https://github.com/psanghal/Springboard-Data-Science/blob/master/Capstone%20Project%201/Project%20Files%20%26%20Data/Capstone1-Final%20Review%20Version-3%20.ipynb

Thank you

Prashant Sanghal